Tentamentsskrivning i **Mathematisk statistik TMA321, 3p.**

Tid: Lördagen den 20 maj, 2006 kl 8.30-12.30
Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736 907 613, MV-huset rum H3026.
Hjälpmedel: valfri räknare, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.

There are five questions with the total number of marks 30. Attempt as many questions, or parts of the questions, as you can. Preliminary grading system (including eventual bonus points):
grade "3" for 12 to 17 marks,
grade "4" for 18 to 23 marks,
grade "5" for 24 and more marks.

---

1. (6 marks) If an infection is present in a school it would be expected to spread to 10% of the children.

a. If you test ten school children at random, what is the probability that the infection is not detected given that it is present?

b. How many children should be tested to have a probability of 0.95 of detecting the infection if it is there?

c. What is the effect of the total number of children in the school on these calculations? Explain.

2. (6 marks) The following data were collected in the end of 19-th century on the number of men killed by a horse in certain Prussian army corps in 20 years, the unit being one army corps (armékår) for 1 year:

Number of deaths	0	1	2	3	4	5 and more
Number of units	144	91	32	11	2	0

a. Why the Poisson distribution might be an appropriate model for the distribution of deaths? (Hint: one soldier - one Bernoulli trial.)

b. For the Poisson parametric model Pois(λ) both the method of moments estimate and the maximum likelihood estimate of λ are equal to the sample mean. Find the MLE of λ using the data.

c. Using the estimate of λ found in (b) compute the probability $\mathrm{P}(X \geq 1)$ for the Poisson distribution. Compare it to the corresponding sample proportion.

d. Estimate the variance σ^2 of the number of deaths in an army corps during one year. Compare it to the estimate of λ found in (b). Why this comparison is meaningful?

3. (6 marks) The following figures show the data from a study of 20 patients with chronic congestive heart failure. Two measurements are shown - ejection fraction, which is a measure of left venticular dysfunction, and pulmonary arterial wedge pressure:

Patient	Ejection fraction (%)	Wedge pressure (mm Hg)
1	28	15
2	26	14
3	42	15
4	29	12
5	16	37
6	21	30
7	25	7
8	35	14
9	30	28
10	36	13
11	37	5
12	41	13
13	20	24
14	26	8
15	38	13
16	26	17
17	10	27
18	18	29
19	10	8
20	31	5

a. Draw a scatterplot for the data. Put ejection fraction on the x-axis and wedge pressure on the y-axis. One value has been mistranscribed from the paper. Which patient's data is most likely to be wrong?

b. Disregarding the erroneous data point, fit by eye an ellipse contour to the scatter plot, assuming that the joint distribution of two factors is bivariate normal. With the help of this ellipse contour draw a regression line relating the wedge pressure to the ejection fraction value.

c. Compare your eye fitting results with the least squares regression line using the next estimates from the data on 19 points: correlation $= -0.64$,

	X	Y
mean	28.2	17.2
standard deviation	8.7	9.3

d. Forty one percent of the variation in wedge pressure can be explained by the variation in ejection fraction. Clarify this statement by referring to a decomposition of the sum of squared deviations for the response variable.

4. (6 marks) It is proposed that the polygraph (lie-detector) be used in association with questioning potential blood donors about whether they are drug users.

There are two kinds of results of the polygraph test: a positive result states that the subject lies, and a negative result says that the subject tells the truth. The polygraph test is known to be 76% sensitive and 63% specific. This means that the conditional probability of a positive result given that the subject lies is 0.76 (the proportion of true positives). On the other hand, the conditional probability of a negative result given that the subject tells the truth is 0.63 (the proportion of true negatives).

In the following we distinguish between potential donors (applicants who claim they do not use drugs) and donors (those who passed the polygraph test).

a. Find the proportion of drug-users claiming they do not use drugs who will be accepted as donors (false negatives) and the proportion of non-users accepted as donors. Assume that all non-users tell the truth.

b. If 6% of applicants use drugs and a third of them lie about it, what is the proportion of drug users among potential donors?

c. What proportion of blood donations will be from drug users?

5. (6 marks) Eight diabetic patients had plasma glucose levels (mmol/l) measured before and one hour after oral administration of 100 g glucose with the following results

Patient	Before	After	Change
1	4.67	5.44	0.77
2	4.97	10.11	5.14
3	5.11	8.49	3.38
4	5.17	6.61	1.44
5	5.33	10.67	5.34
6	6.22	5.67	-0.55
7	6.50	5.78	-0.72
8	7.00	9.89	2.89

a. Estimate the mean change in plasma glucose and compute the standard error of this point estimate.

b. Justify the formula you use for the standard error.

c. Find a 95% confidence interval for the mean change. What assumptions concerning the data do you take in this calculation? Does this assumption look

reasonable judging from the data?

d. What can be said about the P-value of an appropriate test of the null hypothesis $H_0 : \mu_{\text{before}} = \mu_{\text{after}}$ against the alternative $H_1 : \mu_{\text{before}} < \mu_{\text{after}}$?

Statistical tables supplied:
1. Normal distribution table
2. t-distribution table

Good luck!

ANSWERS

1a. Given an infection is present the probability of one randomly chosen child is not infected is 0.9. Assuming independence ten randomly chosen children are all not infected with probability $0.9^{10} = 0.35$.

1b. Assuming independence among n observations, the probability of failing to detect the infection is 0.9^n. With $n = 29$ the probability of detection $1-0.9^n$ becomes smaller than 0.05.

1c. The assumption of independence is relevant if the total number of school children N is much larger than the sample size n. In this sense the previous calculations are approximate. The exact probability of failing to detect the existing infection is

$$\frac{k}{N} \cdot \frac{k-1}{N-1} \cdots \frac{k-n+1}{N-n+1}$$

where $k = 0.9N$ is the number of non-infected children.

2a. Let p be the probability of a soldier to be killed by a horse during a particular year. Assuming independence between n soldiers in an army corp we can model the number of accidents per unit with a $\text{Bin}(n,p)$ distribution. Since n is relatively large and p is small, the $\text{Bin}(n,p)$ distribution can be approximated by a Poisson distribution with parameter $\lambda = np$ which is the accident rate per army corp per year.

2b. The sample mean is $\bar{X} = 0.7$. Thus the MLE of the Poisson parameter λ is $\hat{\lambda} = 0.7$.

2c. With the Poisson model

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda} \approx 1 - e^{-0.7} = 0.50.$$

This probability is close to the corresponding sample proportion $1 - \frac{144}{280} = 0.49$.

2d. The sample variance is

$$s^2 = \frac{n}{n-1}(\overline{X^2} - \bar{X}^2) = 0.763$$

which is rather close to the the sample mean 0.7. This is another evidence in favor of the Poisson model with its variance equal to the mean.

3a. See Figure 1.

3b. After an ellipse contour is sketched, ouline two vertical tangent lines to the ellipse. Then finally draw a straight line passing through the central point

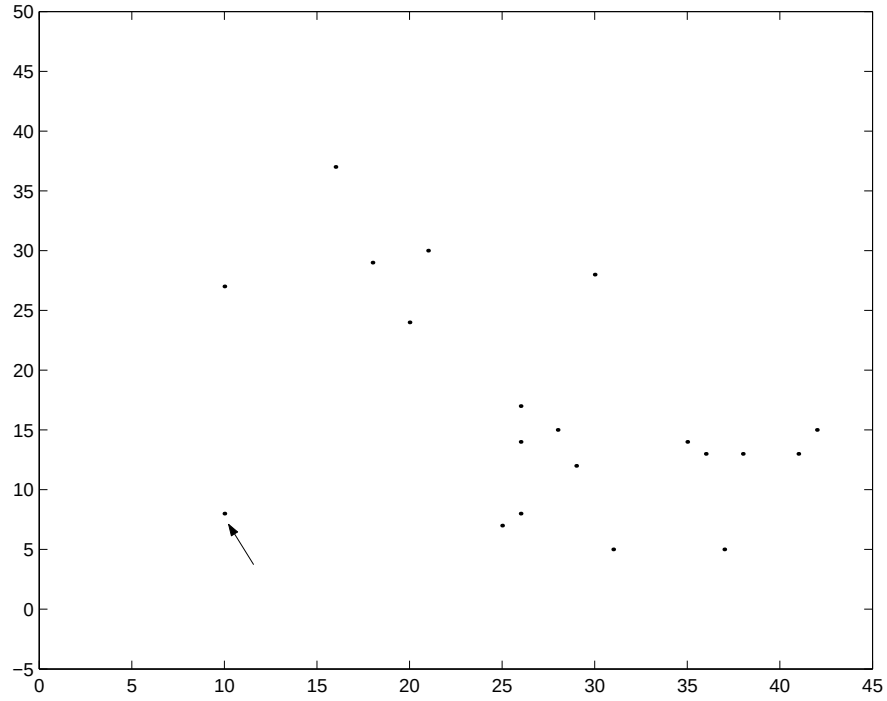

Figure 1: *The scatter plot*

of the data and the two tangent points.

3c. The least squares regression line

$$\hat{y} = 17.2 - 0.64 \cdot \frac{9.3}{8.7}(x - 28.2) = 36.5 - 0.68x.$$

3d. The coefficient of determination $r^2 = (-0.64)^2 = 0.41$ can be expressed as $r^2 = \frac{\text{SSR}}{\text{SST}}$ the proportion of variation in the response variable explained by the explanatory variable. The remaining $\frac{\text{SSE}}{\text{SST}} = 59\%$ is explained by the external factors described by the noise term ϵ in the linear regression model. Illustrate by the diagrams describing the three sums of squares in the decomposition SST=SSR+SSE.

4. Random experiment: pick an individual at random from the population of applicants and test him/her with the polygraph. Random events

A = the person is a drug user, $\bar{A}$ = the person does not use drugs

B = the person lies, $\bar{B}$ = the person tells the truth

C = positive result of the polygraph, $\bar{C}$ = negative result of the polygraph

$D = (A \cap B) \cup \bar{A}$ the person claims not using drugs (potential donor)

Sensitivity $P(C|B) = 0.76$, specificity $P(\bar{C}|\bar{B}) = 0.63$.

4a. The proportion of drug-users claiming they do not use drugs who will be accepted as donors (false negatives) $P(\bar{C}|B) = 1 - 0.76 = 0.24$. The proportion of non-users accepted as donors $P(\bar{C}|\bar{B}) = 0.63$.

4b. If 6% of applicants use drugs and a third of them lie about it, then $P(A) = 0.06$ and $P(B|A) = 0.33$. The proportion of potential donors among all applicants is $P(D) = P(A \cap B) + P(\bar{A}) = 0.02 + 0.94 = 0.96$. The proportion of drug users among potential donors is $P(A|D) = 0.02/0.96 = 0.021$.

4c. Random experiment: pick an individual at random among potential donors and test him/her with the polygraph. Random events

A_1 = the person is a drug user, $\bar{A}_1$ = the person does not use drugs

D_1 the person is accepted as donor

According to 4b $P(A_1) = 0.021$ and $P(\bar{A}_1) = 0.979$. We have also

$$P(D_1|A_1) = 0.24, \ P(D_1|\bar{A}_1) = 0.63.$$

Using the Bayes formula we find the proportion of blood donations from drug users

$$P(A_1|D_1) = \frac{0.021 \cdot 0.24}{0.021 \cdot 0.24 + 0.979 \cdot 0.63} = 0.008.$$

5a. A random sample of size eight $X_1 = 0.77, X_2 = 5.14, X_3 = 3.38, X_4 = 1.44, X_5 = 5.34, X_6 = -0.55, X_7 = -0.72, X_8 = 2.89$. Sample mean $\bar{X} = 2.21$, sample variance $s^2 = 5.58$, sample standard deviation $s = 2.36$, and standard error of the sample mean $s_{\bar{X}} = s/\sqrt{n} = 0.84$.

5b. The standard error of an unbiased point estimate is an estimated standard deviation of the point estimate. The formula $s_{\bar{X}} = s/\sqrt{n}$ comes from the relation

$$Var(\bar{X}) = \frac{Var(X_1) + \ldots + Var(X_n)}{n^2} = \sigma^2/n$$

after σ is replaced by s. Here we used the assumption of independence of n observations taken from the same population distribution with the variance σ^2.

5b. Assume that the changes are normally distributed. According to the t-distribution table with seven degrees of freedom an exact 95% CI of the mean change μ is $2.21 \pm 2.365 \cdot 0.84 = 2.21 \pm 1.97$.

Checking the assumption of normality of the data. There are too few observartions to clearly see a curve behind the $n = 8$ points scattered on the line. Two features supports the assumption of normality: four points are to the right and four points are to the left of the sample mean, there are no obvious outliers lying outside the interval $2.21 \pm 1.96 \cdot 2.36 = 2.21 \pm 4.62$ (notice the use of s

here instead of $s_{\bar{X}}$ as for the CI.)

5c. Test $H_0 : \mu = 0$ against one-sided $H_1 : \mu > 0$. The observed t-test statistic is $T = \frac{2.21}{0.84} = 2.65$. Thus according to the t-distribution table with seven degrees of freedom the P-value of the test lies between 1% and 2.5%.